ONE CONCEPT, TWO FACETS: RETHINKING THE READOUT-VS-CONTROL VERDICT IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear probes, classifiers fit to a model's intermediate representations, decode high-level concepts from residual-stream activations with near-perfect accuracy, yet a decodable direction does not always steer generation. Prior work attributes this gap mainly to properties of the concept, but cannot explain why the same concept yields opposite verdicts under different extraction methods. We show that the readout/control verdict is jointly set by the concept's **structure** and the **extraction pipeline**, and distinguish two poles: a **readout** direction whose activation traces the surface tokens realizing a concept, and a **control variable** direction that shapes downstream generation independently of those tokens. To test this claim, we build a concept bank balanced by concept structure and develop a structure-adaptive protocol that derives the steering layer, budget and coherence threshold from the model under test, scoring effects against a norm-matched random-direction null across five independent model families. Concept structure is positively correlated with steerability in all five families on both splits, and the held-out correlation is significant in every family. Holding the concept and data fixed, selecting attention heads by separability recovers the readout facet while an unselected residual difference-of-means recovers the control facet. Alignment audits that read concepts off such directions may therefore miss what the model acts on.

1 INTRODUCTION

Steering vectors have become a practical tool for controlling large language model (LLM) behavior at inference time: extract a direction from the model's residual stream that encodes a target concept, inject it during generation, and the output shifts accordingly. The directions are found by linear probes, classifiers fit to a model's intermediate activations that test whether a concept is linearly represented. Because probe accuracy is high on a wide range of concepts, an implicit assumption underpins alignment audits, representation engineering, and model editing (Meng et al., 2022a;b; Wang et al., 2024): a linearly decodable direction is also a controllable one. Yet this assumption can fail: on the same corpus, injecting a direction for journalistic stance shifts the model's output from pessimistic to optimistic coverage, while injecting an equally decodable direction for a topical entity changes nothing. What determines whether a decodable direction reads or controls remains open.

A growing body of work has documented this dissociation and proposed explanations, from behavior-type dependence to geometric predictors of steerability (Bas & Novak, 2025; Braun et al., 2025; Gao et al., 2026). However, existing accounts locate the verdict mainly in the concept or the behavior and cannot explain why the same concept yields opposite verdicts under different extraction pipelines.

To address this question, we propose that the verdict is not determined by the concept alone but jointly by the concept's structure and the extraction pipeline, as Fig. 1 illustrates. We distinguish two functional poles within a single concept's representation: a **readout** direction assembled by copying from the concept's surface tokens, and a **control variable** direction written upstream and transportable to unseen inputs. Across five independently trained families, a concept's structure, whether it names a property of the output or a mode of generation, predicts which facet dominates, with no sign reversals.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

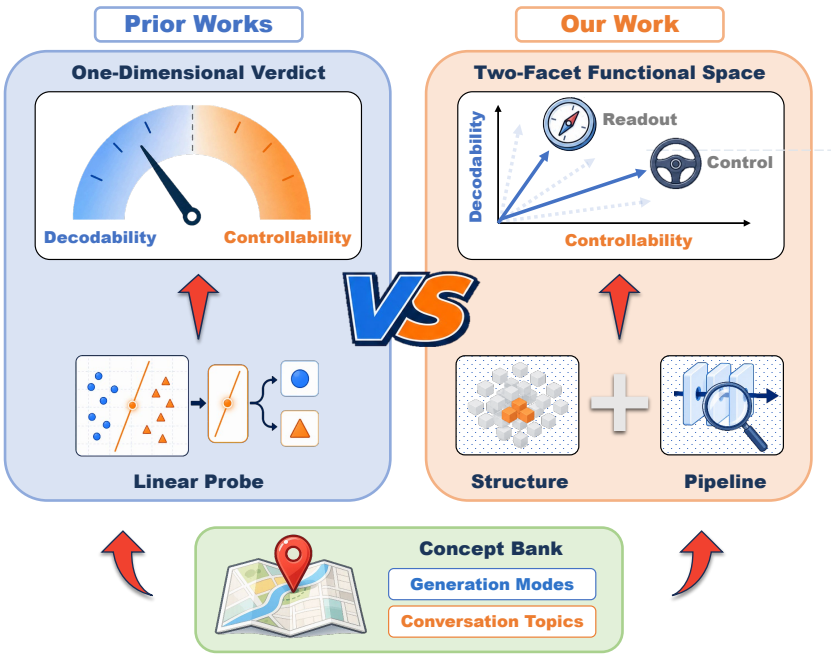


Figure 1: Prior work treats the readout/control verdict as a property of the concept alone. We show it is set jointly by the concept's structure and the extraction pipeline.

The contributions of this paper can be summarized as follows:

- We show that the readout/control verdict for a given concept is not determined by the concept alone but depends jointly on its concept structure and the extraction pipeline, with head-level separability selection and an unselected residual difference-of-means recovering opposite facets of the same concept.
- We propose a structure-adaptive measurement protocol and a type-balanced concept bank that enable fair cross-family comparison of steering behavior.
- We show that on a type-balanced concept bank the leading geometric predictors of steerability, including the logit-lens linear accessibility score, correlate negatively with steerability, and that separability and probe accuracy carry no signal once concept structure is partialled out.
- We provide a causal source-token ablation experiment showing that readout and control directions draw on different information sources.

2 Related Work

2.1 The Decodability-Controllability Gap and Extraction Effects

The observation that a linearly decodable direction need not be controllable has roots in amnesic probing (Elazar et al., 2021) and LEACE (Belrose et al., 2023b), which erase linear information yet leave downstream behavior intact, suggesting the erased directions are read but not used (see also Ravfogel et al., 2022). Nadaf (2026) reports the inverse pole: steerable but not decodable function vectors (Todd et al., 2024). AxBench (Wu et al., 2025) scales the comparison, finding difference-of-means best for detection while prompting beats every steering method, but does not isolate why the same direction fails to steer.

A related question is whether extraction method changes the behavioral outcome. Venkatesh & Kurapath (2026) show that different extractions yield geometrically divergent but behaviorally near-

equivalent directions. We find that head-level separability selection and an unselected residual difference-of-means yield opposite behavioral verdicts on the same concept.

2.2 Determinants and Predictors of Steerability

A second line of work asks what determines whether a direction steers at all, and whether that can be predicted before intervention. Tan et al. (2024) show that steering reliability varies widely across inputs and datasets. Bas & Novak (2025) find efficacy behavior-type-dependent and vector-separation metrics unpredictive of steering (but see Braun et al., 2025), but compare different behaviors where a failure can reflect intrinsic difficulty rather than representational signature. We hold the subject fixed and extract two directions of different concept structure from entity-matched references. Arad et al. (2025) operationalize input/output features (cf. Gur-Arieh et al., 2025) of sparse autoencoders (Cunningham et al., 2023; Gao et al., 2025; Rajamanoharan et al., 2024a;b; He et al., 2025) through the logit lens (nostalgebraist, 2020; Belrose et al., 2023a) and show the steerability asymmetry, but classify features by role; we carry their reading forward to concept directions.

On the geometric side, recent work seeks to predict steerability from representation geometry or dynamics (Park et al., 2024): CRH (Gao et al., 2026), geometric-stability canaries (Raju, 2026), ASTEER (Fan et al., 2026), and the Linear Accessibility Profile (LAP; Billa, 2026). No prior bank is type-balanced, and on ours the leading geometric predictors largely track concept type.

3 Methodology

3.1 Extraction Setup

Let M be a decoder with hidden size n, and write $h_\ell(x,t) \in \mathbb{R}^n$ for the residual stream of input x at layer ℓ and position t. For a response span $T(x)$ define the response-averaged activation

$$\bar{h}_\ell(x) \;=\; \frac{1}{|T(x)|} \sum_{t \in T(x)} h_\ell(x,t). \tag{1}$$

A unit steering direction $\hat{v}_\ell \in \mathbb{R}^n$ enters generation at layer ℓ through the additive hook (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023)

$$h_\ell(x,t) \;\longmapsto\; h_\ell(x,t) + \alpha\, \hat{v}_\ell, \tag{2}$$

applied at every position with multiplier α. Fixing $\|\hat{v}_\ell\|{=}1$ and sweeping α symmetrically holds the perturbation budget constant at a given layer, so a comparison at that layer is between directions only.

Let F and N be concept-framed and neutral continuations of the same n_p prompts, with $|F| = |N| = n_p$. Both pipelines start from the same difference-of-means estimate at layer ℓ,

$$v_\ell \;=\; \frac{1}{|F|} \sum_{x \in F} \bar{h}_\ell(x) \;-\; \frac{1}{|N|} \sum_{x \in N} \bar{h}_\ell(x), \qquad \hat{v}_\ell = \frac{v_\ell}{\|v_\ell\|}. \tag{3}$$

The unselected pipeline takes $\hat{v}_L$ at a fixed layer L. The selected pipeline adds a separability gate. Projecting each reference onto the candidate direction of its own layer gives the scalar $s_\ell(x) = \hat{v}_\ell^\top \bar{h}_\ell(x)$, whose class separation is the signal-detection d',

$$B(\ell) \;=\; \frac{\mu_F(\ell) - \mu_N(\ell)}{\sqrt{\big[\sigma_F^2(\ell) + \sigma_N^2(\ell)\big]/2}} \;=\; \frac{\|v_\ell\|}{\sqrt{\big[\sigma_F^2(\ell) + \sigma_N^2(\ell)\big]/2}}, \tag{4}$$

where $\mu_F(\ell)$ and $\sigma_F(\ell)$ are the mean and standard deviation of s_ℓ over F, and $\mu_N(\ell), \sigma_N(\ell)$ likewise over N. The second equality follows from $\hat{v}_\ell^\top v_\ell = \|v_\ell\|$, so the gate scores a layer by the length of its difference-of-means vector relative to the within-class spread along that vector. The selected direction is extracted at the most separable layer,

$$\hat{v}_{\text{sel}} = \hat{v}_{\ell^*}, \qquad \ell^* = \arg\max_{\ell \in \mathcal{L}} B(\ell), \tag{5}$$

with $\mathcal{L}$ the swept layer set.

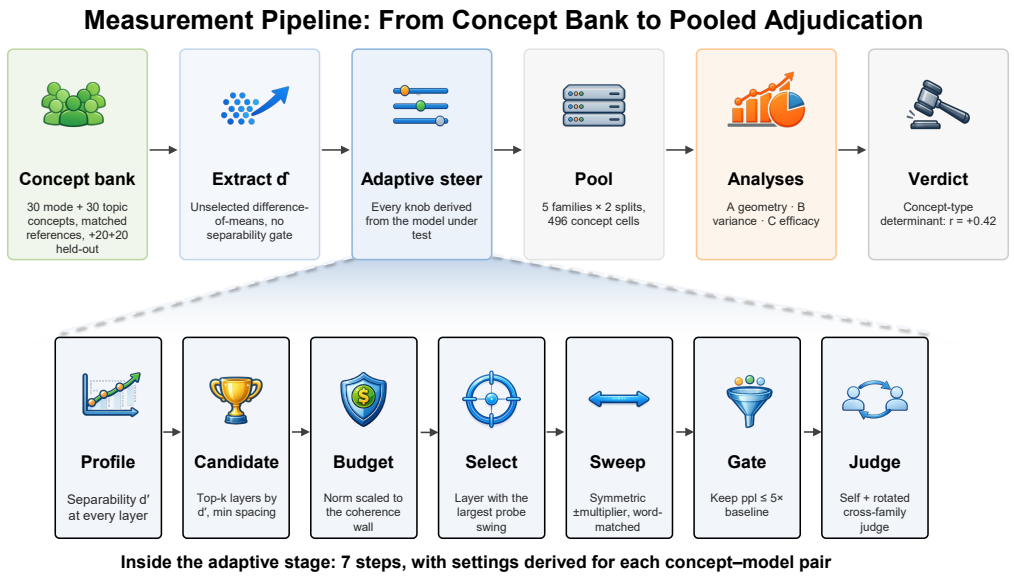


Figure 2: The measurement pipeline, from the type-balanced concept bank to the pooled adjudication, with the structure-adaptive protocol expanded below.

3.2 Diagnostic Battery

Five diagnostics are applied to each extracted direction.

Decodability and separability. A linear probe trained on the direction's scalar projection measures class information as B_{probe} (Marks & Tegmark, 2023), and the separability $B_{d'}$ of Eq. 4 is computed at every layer.

Free-generation steering. The direction is applied through Eq. 2 at a symmetric range of multipliers, and the resulting generations are scored by a continuous judge. A direction that is decodable but produces no swing is functionally a readout.

Cross-context transfer. The direction is extracted then applied to held-out entities never seen during extraction.

Source-token ablation. Content-matched source-token ablation traces where the direction's signal originates, from attended source tokens or from upstream computation (Geiger et al., 2025).

Deterministic constraint checking. For constraint-type concepts, a deterministic verifier checks whether steering changes compliance with numbered constraints.

3.3 Concept Banks

Four stimulus sets are used. A multi-concept zoo spanning affect, register, and topic categories provides breadth, drawn from public classification corpora (Maas et al., 2011; Zhang et al., 2015; Saravia et al., 2018; Pavlick & Tetreault, 2016; Pang & Lee, 2004; Borkan et al., 2019); concepts that conflate class boundaries are pre-registered exclusions. A type-balanced adjudication bank with generation-mode and topic concepts in equal proportion, class known by construction and matched references, supplies the central readout/control contrast. Frame corpora provide transfer tests: both sides name the same entities and differ only in the interpretive lens, so entity vocabulary cancels in Eq. 3 and transfer events use entities disjoint from extraction. A constraint corpus supplies judge-free ground truth: numbered constraints paired with satisfying and violating references, scored by verifiers (in the style of IFEval; Zhou et al., 2023).

Algorithm 1 Structure-Adaptive Steering Protocol

Require: Model $\mathcal{M}$, layers $1{:}L$, contrast (F, N)
Ensure: Steering score St_c
▷ *Profile + Candidates: locate the concept*
for $\ell = 1$ **to** L **do**
  $d'_\ell \leftarrow \mathrm{Sep}(F, N, \ell)$
end for
$\mathcal{C} \leftarrow \{\ell : d'_\ell$ among top-$k\}$
▷ *Budget: calibrate norm to the coherence gate*
$b_c \leftarrow \mathrm{CalibrateBudget}(\mathcal{M})$
▷ *Select: layer that most moves the probe*
$\ell^* \leftarrow$ NIL, $b^* \leftarrow -\infty$
for $\ell \in \mathcal{C}$ **do**
  $b \leftarrow \mathrm{ProbeSwing}(\ell, b_c)$
  if $b > b^*$ **then**
    $\ell^* \leftarrow \ell,\ b^* \leftarrow b$
  end if
end for
▷ *Sweep + Gate: dose-response, keep coherent*
$G \leftarrow \varnothing$
for $m = -m_{\max}$ **to** $m_{\max}$ **do**
  $G_m \leftarrow \mathrm{Steer}(\mathcal{M}, \ell^*, m\, b_c)$
  if $\overline{\mathrm{PPL}}(G_m) \le \mathrm{clip}(5\,\mathrm{PPL}_{\mathrm{base}},\ 8,\ 40)$ **then**
    $G \leftarrow G \cup G_m$
  end if
end for
$m^* \leftarrow \max\{m > 0 : G_{+m}, G_{-m} \subseteq G\}$
▷ *Judge + Verdict: score vs. random null*
$\Delta \leftarrow \mathrm{J}_{\mathrm{self}}(G_{+m^*}) - \mathrm{J}_{\mathrm{self}}(G_{-m^*})$
$\mathrm{St}_c \leftarrow \Delta / (2\, m^* b_c\, \|\mathrm{resid}_{\ell^*}\|)$
return St_c

3.4 Structure-Adaptive Protocol

The steering layer, budget and coherence threshold are derived from the measured model, per concept, as Algorithm 1 sets out and Fig. 2 expands. This places each model at a comparable operating point and replaces an exhaustive sweep over layers and budgets with a short candidate list. Separability d' only shortlists candidate layers, and the steered direction at each layer is the unfiltered difference-of-means of Eq. 3. Every random null direction passes through the same layer selection, budget calibration and coherence gate as the concept direction, so selection alone cannot produce a swing above the null.

3.5 Models

Five independent families of native instruction-tuned dense decoder LMs are used (Yang et al., 2025; Grattafiori et al., 2024; Team et al., 2024; Jiang et al., 2023; Walsh et al., 2025), spanning different training recipes (Ouyang, 2022), tokenizers, and capacity ranges. The case studies and the single-model comparisons run on Qwen3-4B-Instruct-2507; the adjudication then repeats the full protocol independently on every family. Tab. 3 adds three same-recipe capacity ladders. All experiments were run on a single NVIDIA RTX PRO 6000 Blackwell GPU.

4 Experiments

Throughout, the concept-type determinant is the correlation $r(\text{type}, \mathrm{St})$ between concept structure and budget-normalized steerability.

4.1 Experimental Setup

Concept bank. A **mode** concept holds the prompts fixed and varies only the writing manner, for example sarcastic, formal-academic, or melancholic tone. A **topic** concept holds the writing manner fixed and varies what the text is about, for example quantum physics or medieval history. Reference texts are generated once by Qwen3-4B-Instruct-2507 with greedy decoding and shared across all families. The primary bank has 20 modes and 20 topics; the multi-family adjudication expands it to 30+30, and a held-out split of 20 new modes and 20 new topics is disjoint from both.

Judges. In the multi-family adjudication and the zoo, every steered generation receives two scores under the continuous expected-rating protocol of Eq. A.2: a self-score from the steered model itself, and a cross-family score from a judge outside the steered model's family. Concretely, Llama and Gemma judge each other, Mistral and OLMo judge each other, Qwen is judged by Llama, and each ladder rung is judged by whichever of Llama and Gemma lies outside its own family.

4.2 The Two Poles

We first ask whether both poles appear on concrete concepts and extraction alone flips the verdict.

Decodability without control. Two targets are near-perfectly decodable from the residual stream yet move no objective outcome under intervention. Both are scored deterministically, so neither verdict can be a judge artifact.

Constraint satisfaction. An unselected difference-of-means direction between satisfying and violating responses separates the two classes linearly, yet steering with it at matched norm against a random null raises no compliance and at large α only degrades coherence. Because compliance is a deterministic boolean, this is the cleanest demonstration that decodability does not imply control.

Topic as content bias. A probe on the direction "is the text about this subject" separates a single topical entity at ceiling accuracy, yet under free generation the direction installs nothing and instead collapses the prompt's own vocabulary while spiking perplexity. An attention-source decomposition shows that the signal is assembled by attention to the literal entity tokens, far more than for a register control, so the representation exists only because the entity tokens do, which is the defining property of a readout.

The control pole. For the opposite pole we extract an interpretive frame direction as an unselected residual difference-of-means between entity-matched framed and routine references, so the entity cancels and only the generation-mode survives. Steering held-out events with it and sweeping both signs produces a monotone swing, driven mostly by the negative sign, well above a 30-direction random null under both the self-judge and the cross-family judge, as Fig. 3 and Tab. B.2 show. Applied zero-shot to unseen entities absent from extraction it still installs the frame, and the same in-domain and cross-entity transfer replicate on Llama-3.1-8B steered with its own direction and judged by Qwen outside the formal rotation, which confirms a control encoding. Off-domain prompts, where the frame does not apply, stay at the null, so the swing is specific to the frame rather than a generic shift in the judge's score.

Breadth. On the concept zoo, extracted and steered with the same fixed settings, affect and register concepts swing more than topic concepts on the mean, although the two classes overlap, as Fig. 4 and Tab. B.1 show.

Two facets from one concept. Applying both pipelines to the same concept on the same data, separability-gated selection at the attention-head level in the style of ITI (Li et al., 2023), rather than at the layer level of Eq. 5, surfaces the entity-readout heads (cf. Yin & Steinhardt, 2025) and concludes "readout," while the unselected difference-of-means recovers the control facet. It is consistent with a multi-component representation (Elhage et al., 2022), carrying a high-variance readout facet that head-level separability selection picks up and a diffuse control facet that it misses. The same two-facet structure appears on the constraint concept with no LLM judge in the loop, which reconciles our reading with Stolfo et al. (2025).

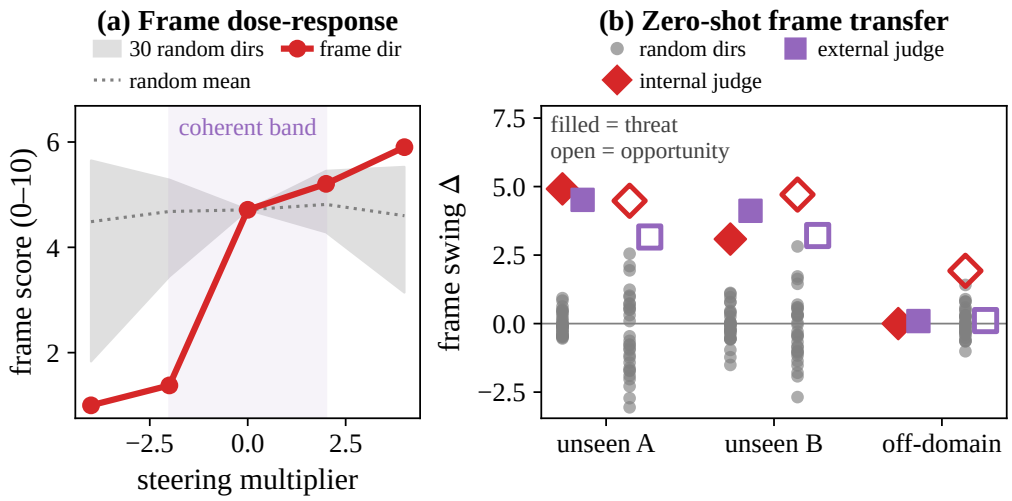


Figure 3: Frame control variable. **(a)** Monotone dose-response, leaving the null band on the negative side. **(b)** Frames transfer zero-shot to unseen entities.

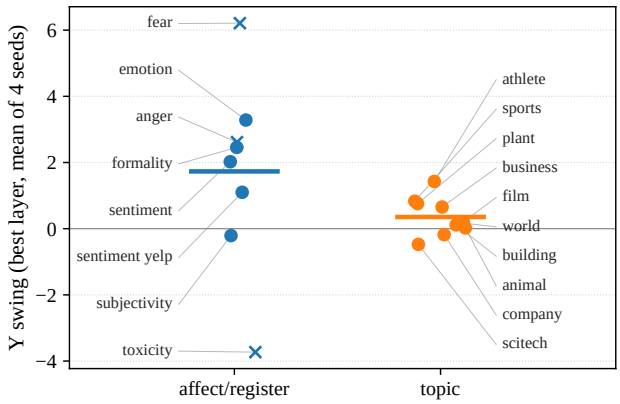


Figure 4: Judged swing on the concept zoo under fixed settings, taken at each concept's best layer and averaged over four seeds. Affect and register concepts swing more than topic concepts on the mean, but the classes overlap. Crosses mark the pre-registered exclusions of Section 3.3.

4.3 Content-Matched Evidence

The two poles come from different concepts, so their contrast could reflect subject matter rather than structure. This test holds subject matter constant and varies only the target's readout/control verdict. For each of five subjects spanning geopolitics, technology, energy, media, and biotechnology, we build two directions from entity-matched references. The content-bias direction contrasts text about S with text about an unrelated subject, an output-property. The stance-mode direction contrasts S framed with a stance against S described routinely.

Both directions receive the identical diagnostic: steer S-absent held-out prompts, score with the continuous judge, compare to a five-direction random null. The dissociation is unanimous across all five subjects: the stance-mode direction is a control variable while the content-bias direction is a readout as Tab. 1 shows.

	Ablation d		Swing z		Swing Δ	
Subject	mode	bias	mode	bias	mode	bias
Geopolitics	-0.42	-6.79	$+3.0$	0.0	$+2.04$	0.00
Technology	$+3.11$	-3.57	$+1.9$	$+1.4$	$+2.05$	$+0.70$
Energy	-0.19	-6.15	$+9.1$	0.0	$+5.13$	0.00
Media	$+6.18$	-2.88	$+5.8$	$+0.1$	$+2.85$	0.00
Biotech	$+2.83$	-5.91	$+7.1$	0.0	$+3.40$	0.00

Table 1: Source-token ablation and matched-pair sweep on the same five subjects. Ablation d = retention in z above a count-matched random-position null; 5/5 subjects, gap z=+7.36. Swing z and Δ = judged swing at ± 2 against a five-direction null with sd floored at 0.5; t=4.6.

4.4 Source-Token Ablation

A readout copies the concept from its own source tokens (Elhage et al., 2021; Ren et al., 2024); severing the copy should collapse it. A control variable is written upstream, away from those tokens; the same lesion should spare it.

On the five matched subjects we mean-ablate attention to S's entity source tokens, specifically the content words of its name and event head-nouns, by redistributing the mass to BOS at layers 0-L. Each direction is re-extracted and scored by how much survives, in z above a count-matched random-position null (Eq. A.5). The same source mask is applied to both directions.

The readout collapses far below the null while the control does not, and the within-subject gap is positive on all five subjects as Tab. 1 shows. The same sign holds on the four families we instrumented, Qwen, Llama, Mistral and OLMo, at their own extraction layers (Qwen and Llama at layer 18, Mistral and OLMo at layer 16).

4.5 Concept Structure vs. Geometry

The case studies above use fixed per-experiment settings, and from here on every result uses the structure-adaptive protocol on the type-balanced bank. If concept structure sets the verdict, it should predict steerability better than geometric quantities that ignore it. We compare concept structure against separability d', probe accuracy, and the logit-lens linear accessibility score A_{lin} of Billa (2026), each correlated with budget-normalized steerability under the structure-adaptive protocol. On Qwen3-4B-Instruct-2507 concept structure is the only positive predictor, with r=+0.41 on the main bank and r=+0.57 on the held-out split, while every geometric quantity correlates negatively, as Tab. B.3 shows. Separability and probe accuracy reduce to concept type, and pooled over all five families their partial correlations with steerability given concept structure are both -0.03. LAP's A_{lin}, implemented on our type-balanced bank, anti-predicts steerability. Given concept structure, its partial r is -0.10 and $+0.01$ on the two Qwen3-4B-Instruct-2507 splits and -0.18 pooled over all five families, so it keeps a weak negative signal beyond concept type. The concept-type confound we identify may equally affect the other geometric predictors of steerability (Raju, 2026; Gao et al., 2026; Fan et al., 2026), all of which are evaluated on unbalanced banks.

4.6 Five-Family Adjudication

We adjudicate on the balanced 60-concept bank across all five families, every protocol knob derived per model and per concept, generations scored by both the self-judge and the rotated cross-family judge. Every family is positive in sign on both splits (Table 2). Pooled over all ten family×split cells with steerability z-scored within each cell, the determinant is r=+0.42 over 496 concept×family pairs, with permutation $p<10^{-4}$ when labels are shuffled within each cell, and averaging each concept across families gives r=+0.63 over 100 concepts. Self- and cross-family judges agree, and the type-beyond-geometry partial stays positive in every cell.

Family	r main [95% CI]	r held-out [95% CI]	Reg.	Agree
Llama-3.1-8B	$+0.43$ [+0.20,+0.64]	$+0.73$ [+0.57,+0.86]	6/10	$+0.86$
Gemma-2-9B	$+0.47$ [+0.30,+0.62]	$+0.50$ [+0.29,+0.70]	8/10	$+0.85$
Qwen3-4B	$+0.41$ [+0.22,+0.57]	$+0.57$ [+0.39,+0.76]	4/10	$+0.80$
OLMo-2-7B	$+0.23^{\dagger}$ [-0.02,+0.44]	$+0.55$ [+0.41,+0.72]	6/10	$+0.84$
Mistral-7B-v0.2	$+0.17^{\dagger}$ [-0.09,+0.34]	$+0.32$ [+0.07,+0.53]	3/10	$+0.78$

Table 2: Five-family adjudication under the structure-adaptive protocol. Four of the $5\times100=500$ concept$\times$family combinations were excluded by the coherence gate. "Reg." means formal registers steerable at $z>2$ out of 10, "Agree" means self/cross-family judge agreement, and † means the main-split CI includes 0.

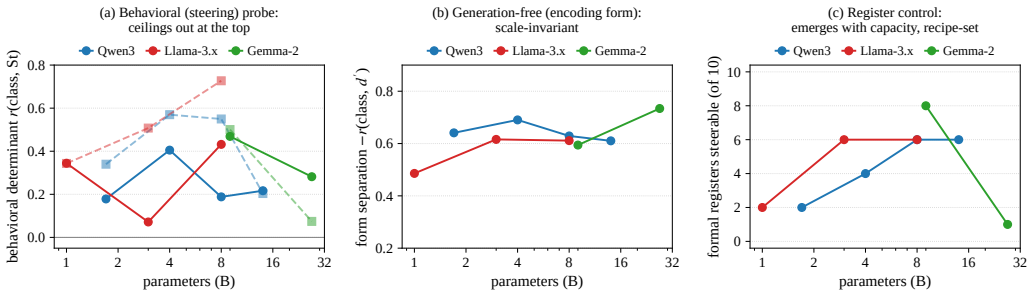


Figure 5: Behavioral determinant across capacity (solid: primary bank; dashed: held-out).

Lineage	Size	r main	r held-out	Reg.	Modes $z>2$
Qwen3	1.7B	$+0.18$	$+0.34^{*}$	2/10	14/30
Qwen3	4B	$+0.41^{*}$	$+0.57^{*}$	4/10	19/30
Qwen3	8B	$+0.19$	$+0.55^{*}$	6/10	17/30
Qwen3	14B	$+0.22$	$+0.20$	6/10	18/28
Llama-3.x	1B	$+0.34^{*}$	$+0.34^{*}$	2/10	13/30
Llama-3.x	3B	$+0.07$	$+0.51^{*}$	6/10	19/28
Llama-3.x	8B	$+0.43^{*}$	$+0.73^{*}$	6/10	21/30
Gemma-2	9B	$+0.47^{*}$	$+0.50^{*}$	8/10	24/30
Gemma-2	27B	$+0.28^{*}$	$+0.07$	1/10	7/25

Table 3: Capacity ladders under identical adaptive protocol. * perm. $p<0.05$; registers out of 10.

4.7 Capacity

We hold the training recipe fixed and vary only capacity (Biderman et al., 2023), as Table 3 and Fig. 5 report. At the smallest rungs the held-out determinant is already significant, and register control appears only at larger rungs. The determinant weakens at the largest rungs, where neither Qwen3-14B split is significant and the Gemma-2-27B held-out correlation falls to $+0.07$.

5 Conclusion

We compare two extraction pipelines on the same data and show that the readout/control verdict is set jointly by the pipeline and concept structure. Content-matched source-token ablation shows that readout and control directions draw on different information sources. The structure effect holds on held-out concepts across five model families. Once concept structure is partialled out, separability and probe accuracy lose their signal, and selecting heads by separability recovers the readout facet. A probe-based audit may therefore measure what a model exposes rather than what it acts on.

AI USE STATEMENT

We used Anthropic's Claude models through the Claude Code command line tool for four tasks. These were LaTeX formatting, polishing the English wording of a draft written by the authors without changing its technical content, helping write the plotting scripts for the figures, and helping write scripts that sweep intervention strength and perplexity thresholds during calibration. The evaluation criteria and acceptance thresholds for that calibration were set by the authors, and the final parameters follow the deterministic rules of Appendix B. Claude was also used outside the measurement protocol to read steered generations and confirm that they remained legible. This check produces no reported number and was not used to filter any data. Generative AI was not used to choose the research question, design the method, run the statistical analysis, interpret the results, or draw conclusions. The open-source models that score generations inside the protocol are part of the method and are described in Section 4.1. The authors reviewed all AI-assisted code and text and take full responsibility for the content of this paper.

ETHICS STATEMENT

All corpora used in this work are public classification datasets, and the reference texts of the concept bank were generated by an open-weight model. The data contain no personal information and the study involves no human subjects. Activation steering can change model behavior at inference time. Our finding that separability-based evaluations report the readout facet is intended to make alignment audits that rely on such directions more reliable.

REPRODUCIBILITY STATEMENT

The extraction pipelines are defined in Section 3.1, the structure-adaptive protocol in Algorithm 1, and every threshold, budget range, layer-selection rule and null construction in Appendices A and B. All five model families are publicly released instruction-tuned checkpoints, reference texts are produced by greedy decoding, and all experiments fit on a single GPU.

REFERENCES

Dana Arad, Aaron Mueller, and Yonatan Belinkov. Saes are good for steering–if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10252–10270, 2025.

Tetiana Bas and Krystian Novak. What can we actually steer? a multi-behavior study of activation control. *arXiv preprint arXiv:2511.18284*, 2025.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023a.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023b.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430, 2023.

Jayadev Billa. Predicting where steering vectors succeed. *arXiv preprint arXiv:2604.15557*, 2026.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un) reliability of steering vectors in language models. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Chenrui Fan, Yize Cheng, Ming Li, Soheil Feizi, and Tianyi Zhou. When is your llm steerable? *arXiv preprint arXiv:2606.11599*, 2026.

Lang Gao, Jinghui Zhang, Wei Liu, Fengxian Ji, Chenxi Wang, Zirui Song, Akash Ghosh, Youssef Mohamed, Preslav Nakov, and Xiuying Chen. The cylindrical representation hypothesis for language model steering. *arXiv preprint arXiv:2605.01844*, 2026.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yoav Gur-Arieh, Roy Mayan, Chen Agassy, Atticus Geiger, and Mor Geva. Enhancing automated interpretability with output-centric feature descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5757–5778, 2025.

Zirui He, Haiyan Zhao, Yiran Qiao, Fan Yang, Ali Payani, Jing Ma, and Mengnan Du. Saif: A sparse autoencoder framework for interpreting and steering instruction following of language models. *arXiv preprint arXiv:2502.11356*, 2025.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in neural information processing systems*, 2022a.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The eleventh international conference on learning representations*, 2022b.

Mohammed Suhail B Nadaf. Steerable but not decodable: Function vectors operate beyond the logit lens. *arXiv preprint arXiv:2604.02608*, 2026.

nostalgebraist. interpreting GPT: the logit lens. LessWrong, https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, August 2020.

L Ouyang. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730, 2022.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 271–278, 2004.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 39643–39666, 2024.

Ellie Pavlick and Joel Tetreault. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74, 2016.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, Janos Kramar, Rohin Shah, and Neel Nanda. Improving sparse decomposition of language model activations with gated sparse autoencoders. *Advances in Neural Information Processing Systems*, 37: 775–818, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024b.

Prashant C Raju. The geometric canary: Predicting steerability and detecting drift via representational stability. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. *arXiv preprint arXiv:2201.12091*, 2022.

Jie Ren, Qipeng Guo, Hang Yan, Dongrui Liu, Quanshi Zhang, Xipeng Qiu, and Dahua Lin. Identifying semantic induction heads to understand in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6916–6932, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 3687–3697, 2018.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *International Conference on Learning Representations*, volume 2025, pp. 55790–55823, 2025.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. *arXiv preprint arXiv:2602.06801*, 2026.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, et al. 2 olmo 2 furious (colm's version). In *Second Conference on Language Modeling*, 2025.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37, 2024.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning*, pp. 67035–67080, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? In *International Conference on Machine Learning*, pp. 72428–72461, 2025.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Formal Definitions

Coherence gate, Alg. 1 line 16. Let G_m denote the steered generations at multiplier m, one per evaluation prompt. They pass the gate iff their mean perplexity satisfies $\overline{\mathrm{PPL}}(G_m) \le \tau_M$, with the model-relative threshold

$$\tau_M \;=\; \mathrm{clip}\big(5 \cdot \mathrm{PPL}_M^0,\, 8,\, 40\big), \tag{A.1}$$

where perplexity is measured under the steered model itself, clip bounds its argument to $[8, 40]$, and PPL_M^0 is model M's baseline perplexity. The multiplier 5 and the clip bounds were calibrated once on pilot generations and then held fixed across all models and concepts.

Judge and swing, Alg. 1 line 21. A generation is scored not by a hard label but by an expected rating. The judge is asked to rate the answer a against the concept question q on an integer scale from 0 to 10, and the rating is read off the single-token distribution rather than from the sampled token,

$$J(a, q) = \sum_r r\, p(r \mid a, q), \qquad p(r \mid a, q) = \frac{\exp \lambda_r}{\sum_{r'} \exp \lambda_{r'}}, \tag{A.2}$$

where r ranges over the integers $\{0, \dots, 10\}$ and the logits λ_r are renormalized over those eleven tokens alone. Reading the expectation makes the score continuous, so a direction that shifts behavior slightly is not rounded away.

The swing of a direction is the judged difference between its two signs at the same magnitude,

$$\Delta(m) = \frac{1}{|S|} \sum_{s \in S} \big[J(g_{+m}(s), q) - J(g_{-m}(s), q)\big], \tag{A.3}$$

with S the evaluation prompt set, m the multiplier grid index, and $g_{\pm m}(s)$ the generations from prompt s under $\pm m$.

Budget-normalized steerability.

$$\mathrm{St}_c \;=\; \frac{\Delta_c(m^*)}{2\, m^* b_c\, \|\mathrm{resid}_{\ell^*}\|}, \tag{A.4}$$

where m^* is the largest grid multiplier whose two signs both pass the coherence gate, Δ_c is Eq. A.3 for concept c under the self-judge, $b_c\|\mathrm{resid}_{\ell^*}\|$ is the norm of the injected direction, with b_c the calibrated budget fraction and $\|\mathrm{resid}_{\ell^*}\|$ the mean residual-stream norm at the steering layer, and $2m^*$ is the span from $-m^*$ to $+m^*$. The cross-family judge scores the same generations and enters only the agreement statistic.

$$z \;=\; \frac{\Delta_{\text{dir}}(m) - \overline{\Delta_{\text{rand}}(m)}}{\mathrm{sd}_k\, \Delta_{u_k}(m)}. \tag{A.5}$$

Here $u_1, \dots, u_K$ are unit directions drawn uniformly from the sphere in $\mathbb{R}^n$ at the same m, and sd_k is taken over them.

B Protocol Details

Multi-family adjudication protocol. These are the settings behind §3.4, used for the geometry comparison §4.5, the five-family adjudication §4.6 and the capacity ladders §4.7. The case studies and single-model races in §4.2-§4.4 predate this protocol and use fixed per-experiment settings. For budget calibration the steering magnitude starts at 0.15 of the mean residual-stream norm at the chosen layer and is scaled until the largest multiplier on the grid $\{-4, -2, 0, +2, +4\}$ sits just inside the coherence gate of Eq. A.1. The final fraction is confined to $[0.05, 0.45]$. For layer selection the separability profile of Algorithm 1 is computed at every layer, and its candidate set $\mathcal{C}$ keeps the four highest-d' layers subject to a minimum spacing of $\max(2, \lfloor n_{\text{layers}}/12 \rfloor)$, so k=4; the steering layer is the candidate whose probe-swing at that budget, the amount by which steering moves the probe, is largest. Because tokenizers differ in granularity, generation length and the pooling window are rescaled by each model's median tokens-per-word ratio. Decoding is greedy throughout; the random

Concept	Type	$B_{d'}$	B_{probe}	Δ self	Δ cross
anger	mode	2.82	0.94	$+2.61$	$+3.00$
emotion	mode	2.24	0.84	$+3.28$	$+2.49$
fear	mode	2.83	0.92	$+6.21$	$+4.45$
formality	mode	3.59	0.96	$+2.46$	$+2.07$
sentiment	mode	2.30	0.87	$+2.02$	$+2.68$
sentiment_yelp	mode	2.06	0.81	$+1.10$	$+3.35$
subjectivity	mode	2.02	0.83	-0.21	-0.07
toxicity	mode	2.32	0.89	-3.73	-0.79
animal	topic	4.57	0.97	$+0.02$	$+0.45$
athlete	topic	3.18	0.93	$+1.43$	$+1.97$
building	topic	3.37	0.92	$+0.11$	$+0.16$
business	topic	2.65	0.92	$+0.66$	$+1.03$
company	topic	3.85	0.97	-0.18	$+0.82$
film	topic	5.15	0.98	$+0.20$	$+0.39$
plant	topic	3.54	0.93	$+0.76$	$+2.45$
scitech	topic	2.55	0.87	-0.47	$+0.98$
sports	topic	3.97	0.96	$+0.83$	-0.65
world	topic	2.46	0.90	$+0.18$	$+0.40$

Table B.1: Per-concept zoo: separability versus the functional swing Δ under both judges. The classes separate on the mean, not as a clean partition.

Setting	Δ	Null $\mu \pm \sigma$	z
In-domain			
Threat (Q+Q)	$+3.83$	$+0.13 \pm 0.54$	6.9
Threat (Q+L)	$+2.99$	$+0.03 \pm 0.31$	9.7
Threat (L+Q)	$+3.68$	-0.34 ± 1.17	3.4
Unseen entities, zero-shot			
Threat A (Q+Q)	$+4.92$	-0.01 ± 0.41	12.0
Threat B (Q+Q)	$+3.08$	-0.05 ± 0.63	5.0
Threat A (Q+L)	$+4.51$	$+0.05 \pm 0.65$	6.9
Threat B (Q+L)	$+4.12$	-0.07 ± 0.64	6.5
Threat A (L+Q)	$+2.54$	$+0.03 \pm 0.42$	6.0
Threat B (L+Q)	$+3.50$	-0.25 ± 0.14	26.9
Opportunity A (Q+Q)	$+4.48$	-0.40 ± 1.46	3.3
Opportunity B (Q+Q)	$+4.71$	-0.21 ± 1.20	4.1

Table B.2: Frame steering and transfer. Δ is the judged swing at fixed ± 2, scored against a norm-matched random-direction null, and z is that swing above the null. Each row names the steered model and the judge in that order, with Q for Qwen3-4B-Instruct-2507 and L for Llama-3.1-8B, so Q+Q is a self-judge and Q+L a cross-family judge.

Predictor	Qwen main	Qwen held-out	Pooled r	Pooled partial
Concept structure	$+0.41$	$+0.57$	$+0.42$	–
$B_{d'}$	-0.31	-0.38	-0.31	-0.03
B_{probe}	-0.25	-0.15	-0.28	-0.03
A_{lin}	-0.34	-0.38	-0.39	-0.18

Table B.3: Geometric predictors against budget-normalized steerability under the structure-adaptive protocol. The Qwen columns report Pearson r on Qwen3-4B-Instruct-2507. The pooled columns report r and the partial correlation given concept structure over all five families, with every quantity z-scored within each family$\times$split cell.

null comprises five norm-matched directions per concept, each drawn independently across four seeds (20 direction$\times$seed samples in total). Each null direction passes through the same adaptive protocol as the concept direction.